Luci e ombre dei *big data*

Antonio Addis, Alessandro Rosa
Dipartimento di Epidemiologia del Servizio Sanitario Regionale, Regione Lazio, Roma

Se sommiamo la quantità di dati che ogni nostra azione produce giornalmente ci rendiamo conto dell'enorme volume di byte che si accumulano intorno a noi. È forse anche per questo che non possiamo parlare più di dati, bensì di *big data*. Secondo alcuni è solo un'operazione di marketing ma molti intravedono invece un nuovo paradigma che dominerà e regolerà l'esistenza di molti.
La stampa popolare e accademica ha iniziato a utilizzare il termine "big data" per descrivere la rapida integrazione e analisi su larga scala; tuttavia, una chiara definizione di *big data* rimane sfuggente [1]. Le modalità con le quali i *big data* potrebbero influenzare il futuro della ricerca epidemiologica e degli interventi sanitari sulla popolazione sono anch'esse, al momento, poco chiare.
In generale ci si è trovati d'accordo sul fatto che il fenomeno possa essere principalmente descritto dalle cosiddette 3 V: volume, varietà e velocità.
Il *volume*: termine percepito per primo fa esclusivamente riferimento alla mole di dati raccolti e dei record che popolano i tanti *data sets* in cui finiscono i nostri dati sanitari. Tecnicamente, anche le schede di dimissione ospedaliera dovrebbero appartenere ai *big data* in quanto composte da milioni di record.
La *varietà*: i dati possono presentare eterogeneità nel tipo, nella rappresentazione e nell'interpretazione semantica. Possono essere di qualsiasi natura (strutturati, semi-strutturati o non strutturati). In questo caso dovremmo pensare comunque alla possibilità di un ipotetico *linkage* tra sistemi informativi riguardanti ad esempio la farmaceutica, i ricoveri, l'assistenza specialistica ecc. per poter parlare di *big data*.
La *velocità*: alle nuove informazioni estraibili dai dati viene spesso associata una funzione di utilità che degrada velocemente con il passare del tempo. La velocità inoltre è anche relativa al tasso di produzione dei dati.
Uno degli elementi che ci fa distinguere quello di cui stiamo parlando dai numerosi dati presenti nei nostri computer dovrebbe essere che i *big data* vengono generati automaticamente da operazioni di interazione persona-macchina (un esempio, in ambito finanziario, sono i dati transazionali), persona-persona (social network) e macchina-macchina (si pensi ai dati inviati dai sensori direttamente ai telefoni cellulari). Nella convenzione universalmente accettata si associano a enormi moli di volume: si passa dai terabyte (1 tb = 1012 b) e petabyte (1 pb = 1015 b), fino ad arrivare agli exabyte e addirittura agli zettabyte. Devono presentare un tasso di produzione alto e, inoltre, possono essere di provenienza varia e talvolta non convenzionale: parliamo anche di documenti testuali, immagini, audio, video, dati da sensori o Gps.
I *big data*, in sintesi, presentano congiuntamente le tre caratteristiche sopra elencate e sono la materializzazione dell'*internet of things*, cioè la visione secondo cui gli oggetti nel mondo informatizzato creano un sistema pervasivo e interconnesso avvalendosi di molteplici tecnologie di comunicazione. In pratica, parliamo di dati e flussi continui [1].
Esiste poi un approccio "attivo", che utilizza la rete per reclutare volontari a cui chiedere informazioni circa la loro condizione di salute. Si tratta sempre di dati digitali generati tramite web ma appositamente per scopi epidemiologici. L'approccio "attivo" è quello su cui si fonda, per esempio, InfluenzaNet, una piattaforma web interattiva volta a raccogliere dati sull'influenza stagionale – con una risoluzione geografica e temporale molto alta – per informare modelli predittivi. La sorveglianza viene condotta su una coorte di volontari che annualmente, all'inizio della stagione influenzale, vengono invitati a riportare la loro condizione di salute sia che stiano bene sia che abbiano dei sintomi respiratori. Con questo approccio non si raggiungono le dimensioni dei *big data*, poiché il numero degli individui raggiunti con questa modalità non è paragonabile ai milioni di utenti di Facebook o Twitter, ma il numero è tale per cui il segnale epidemiologico che si ottiene è sufficientemente accurato. Inoltre, con la modalità della sorveglianza partecipativa si possono ottenere informazioni da persone che non si recano dal medico in caso di febbre, ma che non hanno problemi a compilare un questionario sul web quando sono a casa da malati.
InfluenzaNet è stato sperimentato per la prima volta in Olanda e in Belgio nella stagione influenzale 2003/2004. Ora viene utilizzato in 10 Paesi europei, tra cui l'Italia con Fondazione Isi e l'ISS, la Francia con l'Inserm e l'Inghilterra con la Public Health England, e ha inoltre ispirato delle piattaforme analoghe negli Stati Uniti e in Australia. Si è quindi creato un sensore digitale globale di volontari, sia dell'emisfero nord che di quello sud, che ogni anno durante la stagione influenzale riportano il proprio stato di salute. Questo è un enorme passo avanti nella sorveglianza globale dell'influenza [1]. Tuttavia, è proprio in questo ambito che si è registrato il primo grande flop dei *big data*. L'attività dei motori di ricerca, quale Google che conta centinaia di milioni di utenti attivi, è stata considerata per un certo periodo come un segnale affidabile ma non sempre preciso. *Google flu trends* si basava sull'analisi delle ricerche fatte tramite il motore di ricerca di Google di parole collegate ai sintomi influenzali quali febbre, mal di gola, raffreddore. Il numero di volte che gli utenti chiedevano al motore di ricerca queste informazioni veniva utilizzato come specchio del numero di casi di influenza fra la popolazione. Dopo diversi inverni di mappature perfette delle epidemie influenzali, nel 2013 il sistema ha fallito clamorosamente sovrastimando i casi di influenza. Il problema è che Google usava un modello statistico impiegato per produrre previsioni da una settimana all'altra e che veniva allenato soltanto sui dati della stagione corrente, quando invece la dinamica dell'influenza è tale per cui si osserva sempre lo stesso andamento stagionale, ma se si analizza nel dettaglio si osserva che ogni stagione è diversa. Inoltre, il fatto che una persona cerchi la parola "influenza" con Google non è indicativo del motivo per cui lo fa: potrebbe eseguire la ricerca perché ha l'influenza ma anche perché ne ha sentito parlare molto dai media. In pratica il flop di *Google flu trends* potrebbe essere imputabile non tanto alla qualità dei dati digitali quanto piuttosto al modello di calcolo impiegato che non è mai stato reso noto alla comunità scientifica. La lezione è stata importante per far capire insieme alle potenzialità anche i limiti di questi sistemi [2].

Vi sono inoltre molti aspetti che riguardano il corretto utilizzo di tutti questi dati e che hanno a che fare con il tema della privacy e la qualità del dato raccolto. La messa insieme di tanti dati, per quanto differenziati e in tempi molto veloci ma scorretti, non produrrà di per sé un dato qualitativamente migliore. Insomma anche in questo caso si tratta del fenomeno *garbage in - garbage out*.
Le possibilità di connettere sistemi diversi, combinare dati attraverso linguaggi condivisi, trasformare il rumore di fondo in nuove e utili informazioni, aumentare i punti di osservazione sui fenomeni, rendere più efficienti in termini di tempo e spazio le rilevazioni, sono tutti potenziali vantaggi che sembrano ora possibili. All'analisi di tutti questi aspetti e altri ancora è stato dedicaato uno degli approfondimenti del progetto Forward (http://forward.recentiprogressi.it/), un'iniziativa del *Pensiero Scientifico Editore*, e del *Dipartimento di Epidemiologia della Regione Lazio*, che insieme ad alcune aziende private ha avviato una serie di riflessioni su ciò che diventerà importante nel prossimo futuro nell'ambito del settore sanitario.
In conclusione, le criticità, e le potenzialità associati ai *big data* per la salute pubblica sono numerose e rilevanti. Anzitutto la disponibilità di dati in tempo reale consente di monitorare costantemente l'evolvere per esempio di un'epidemia, o, nell'ambito della farmacovigilanza, di identificare segnali relativamente a eventi avversi a farmaci che possono completare la sorveglianza routinaria e la farmacovigilanza, solitamente effettuate attraverso le segnalazioni da parte degli operatori sanitari. Occorre però ancora del tempo per separare il segnale vero dal rumore di fondo e tradurre le informazioni sempre più numerose di cui disponiamo in benessere e salute per i cittadini.

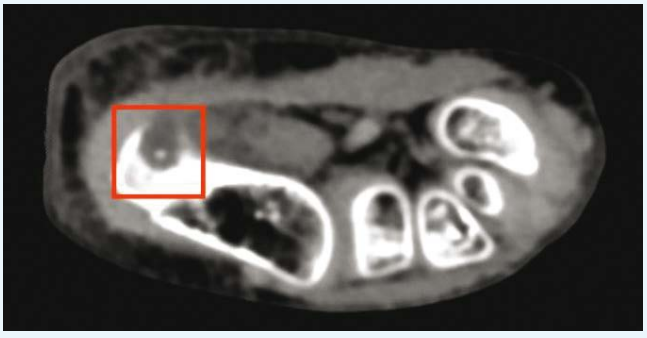
✉ a.addis@deplazio.it

1. Rosa A. Un approccio semantico. *Recenti Prog Med* 2106; Suppl Forward 4;S6-S7. http://forward.recentiprogressi.it/wp-content/uploads/2016/11/suppl4_rosa.pdf
2. Paolotti D, Rizzo C. Le impronte digitali al servizio dell'epidemiologia. *Recenti Prog Med* 2106; Suppl Forward 4;S8-S10. http://forward.recentiprogressi.it/wpcontent/uploads/2016/11/suppl4_paolotti_rizzo.pdf

A COLPO D'OCCHIO

Rubrica a cura di Enrico Valletta e Martina Fornaro

UO di Pediatria, Ospedale G.B. Morgagni - L. Pierantoni, AUSL della Romagna, Forlí



Rx e TAC in bambina con dolore e segni di flogosi al tallone da alcune settimane

Di cosa si tratta?

- ❑ Osteomielite
- ❑ Osteosarcoma
- ❑ Istiocitosi
- ❑ Osteoma osteoide

Soluzione del quesito a p. 126